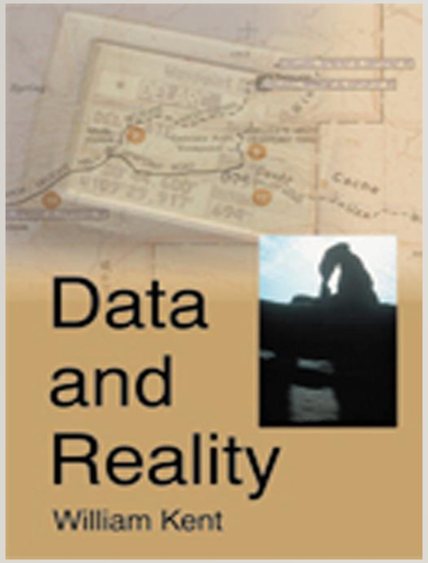
INFORMATION ARCHITECTURE

DATA AND REALITY

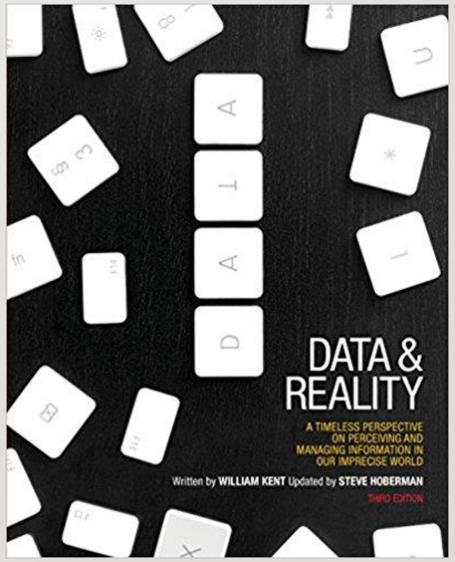
DATA AND REALITY

SECOND EDITION

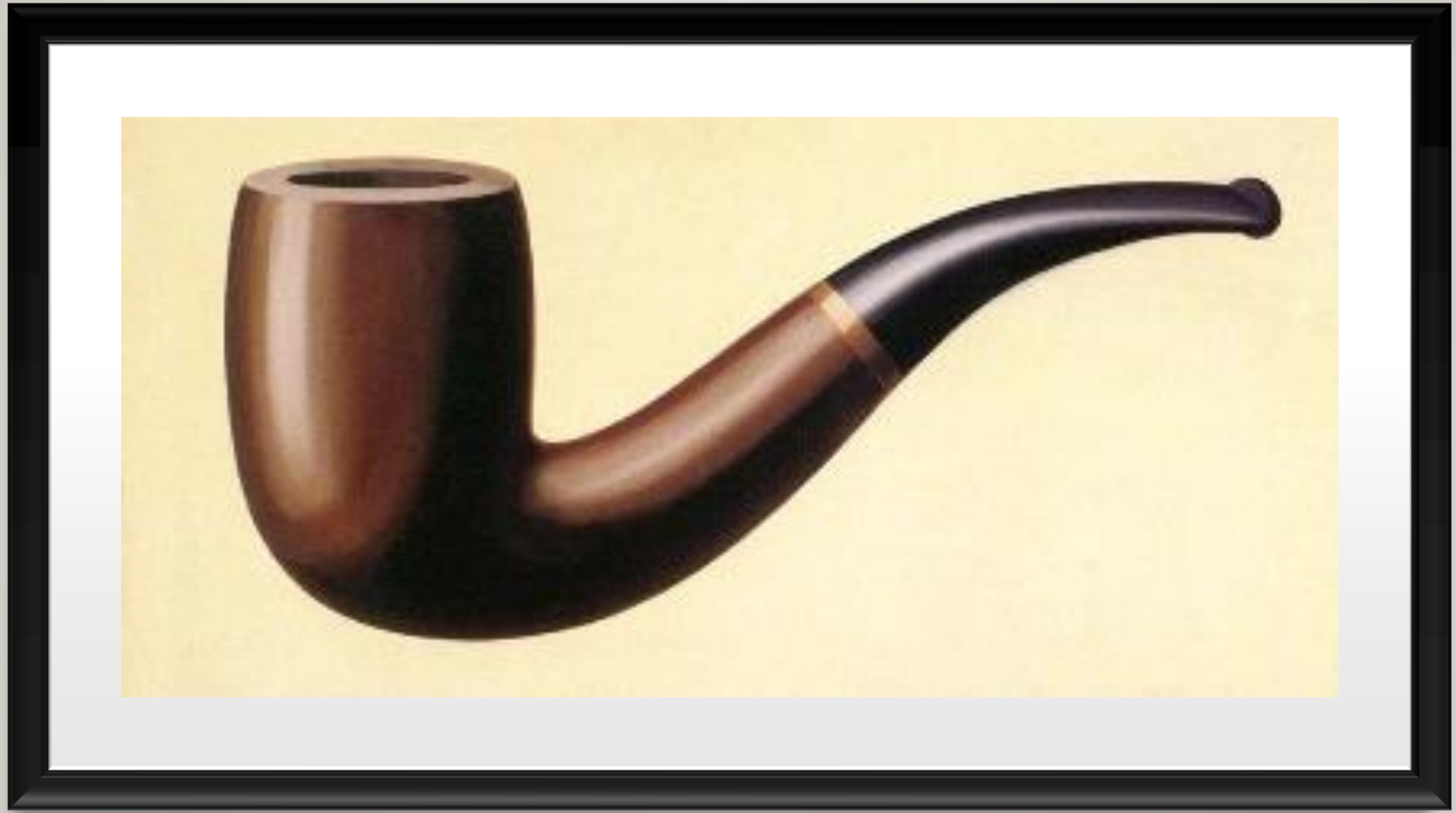


✓

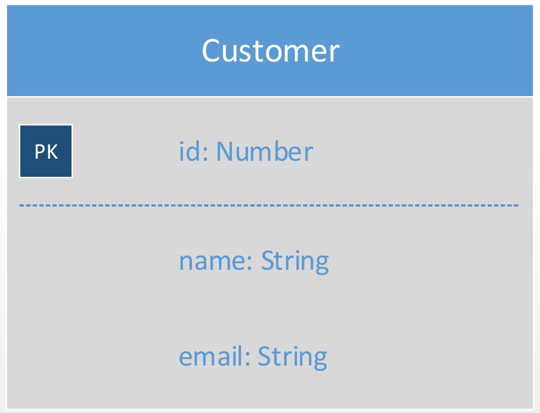
THIRD EDITION


DATA & REALITY
A TIMELESS PERSPECTIVE ON PERCEIVING AND MANAGING INFORMATION IN OUR IMPRECISE WORLD
Written by WILLIAM KENT Updated by STEVE HOBERMAN
THIRD EDITION

✗



"La Trahison des Images" ("The Treachery of Images")
By René Magritte, 1898-1967. The work is now owned by and exhibited at LACMA.



Ce n'est pas un client

THE SYMBOL IS NOT THE THING

- We don’t put people in our databases.
- We put representations of people in our databases.

TYPE:

1. Set of allowed values
2. Operations on those values

TYPE:

1. ~~Set of allowed values~~

   Decision process to determine if a value is a member
2. Operations on those values

ENUMERATED TYPE

We can list all the allowed values

Decision process just checks for membership in the set.

E.g., Boolean or rows of a table

ENCODED TYPE

- Some types are hard to enumerate.
- Very big sets: E.g., all rational numbers.
- Sets subject to redefinition: E.g., Addresses, Names, other social constructs

- We encode these in primitive values

COMMON ENCODINGS

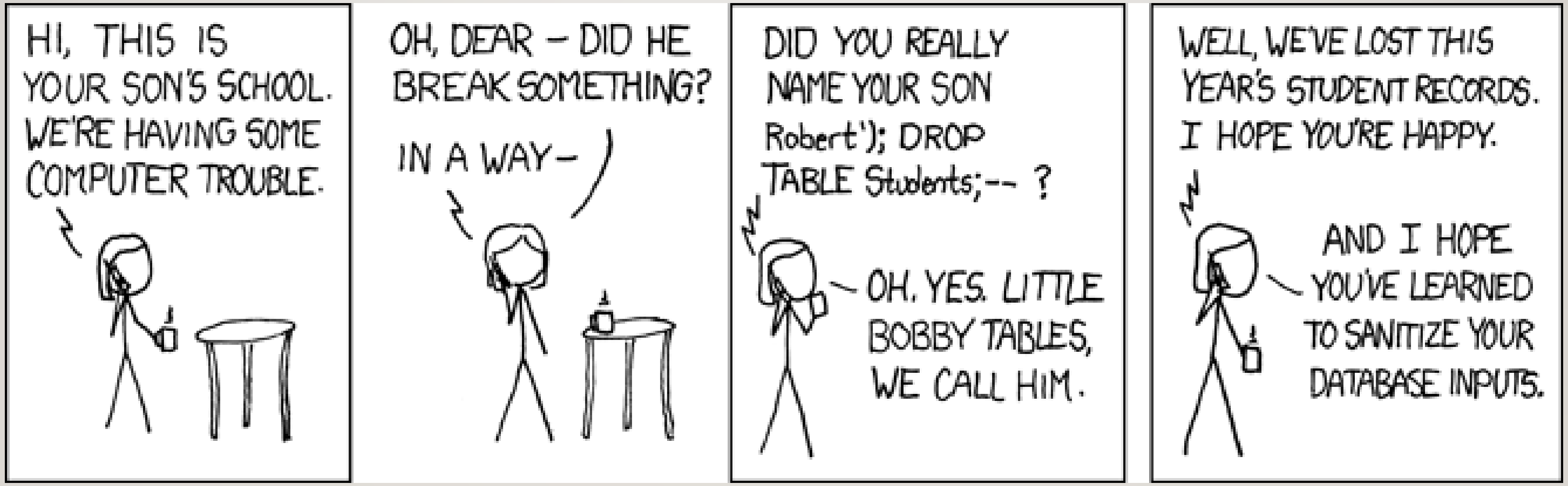
Data type	Representation
Name	String
Height	16-bit 2’s complement
Salary	16-bit 2’s complement

THE TROUBLE WITH ENCODING

- The decision procedure becomes checking syntax
- “Does this value meet the syntax for encoding?”
- Rather than “Is this value a member of this type?”

- Opens door to incorrect acceptance or rejection of values

THIS IS A TYPE ERROR


HI, THIS IS YOUR SON'S SCHOOL. WE'RE HAVING SOME COMPUTER TROUBLE.
OH, DEAR – DID HE BREAK SOMETHING?
IN A WAY–
DID YOU REALLY NAME YOUR SON Robert'); DROP TABLE Students;-- ?
OH. YES. LITTLE BOBBY TABLES, WE CALL HIM.
WELL, WE'VE LOST THIS YEAR'S STUDENT RECORDS. I HOPE YOU'RE HAPPY.
AND I HOPE YOU'VE LEARNED TO SANITIZE YOUR DATABASE INPUTS.

https://www.xkcd.com/327/

ATTRIBUTE TYPES

- Height **isa** Integer?
    - Lacks unit
    - Allows too many operations
    - Allows too many values… what does a negative height mean?

ANOTHER BROKEN ATTRIBUTE TYPE

- Money **isa** double?
  - Imprecise operations, will lose or manufacture money
  - Imprecise representation
  - Lacks currency, so allows illegal conversions and combinations
  - Allows too many operations.
  - What does $\$5\ mod\ 3$ mean?
- Please don’t ever do this.

YET ANOTHER BROKEN ATTRIBUTE TYPE

- Datetime **isa** long?
  - Allows too many operations.
  - Allows incorrect closure under subtraction. $Time - Time \rightarrow Duration$
  - Is it wall clock (goes backward, jumps forward)?
  - Is it system time?
  - What time zone?

PITFALLS OF SYNTAX

Data type	Representati on	Sample value	Pitfall
Name	String	♣♦♥♠↓–↓←→←→™	Canonicalization. Duplicates.
Height	16-bit 2's complement	190	Inches? Meters? Pixels?
Salary	16-bit 2's complement	2.5	'Height' x 'Salary' == ???

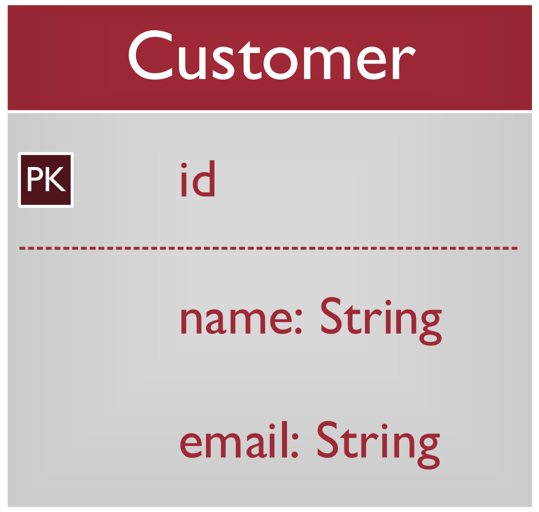
MY RECURRING THEME

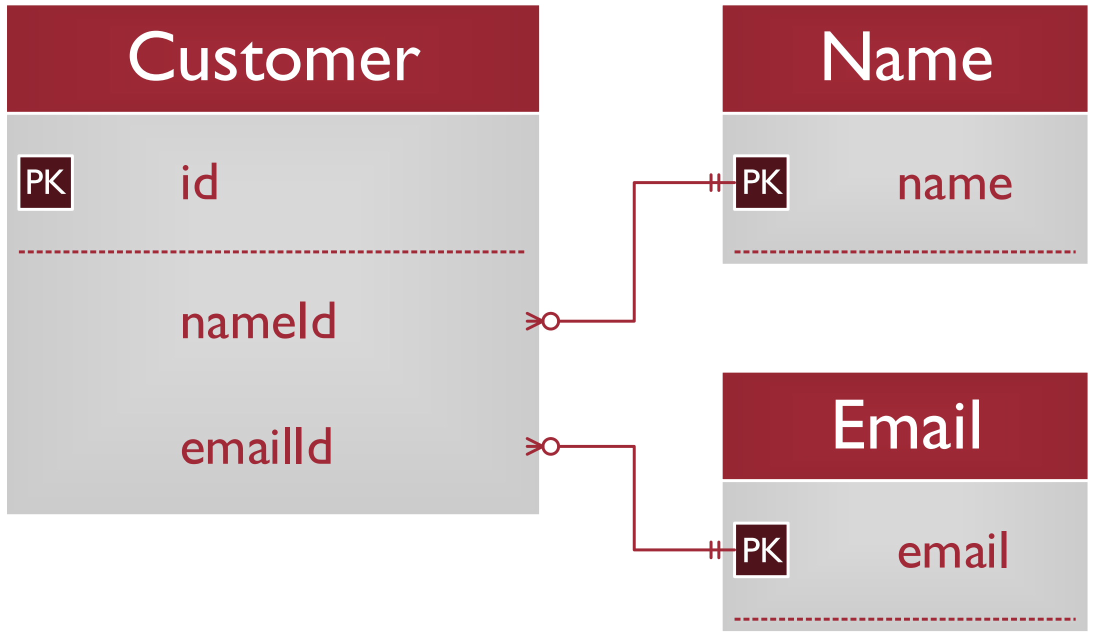
- Focus on the behavior.
- What are valid operations on the attributes?
- Can you make a useful type equation about the attributes?

✗ E.g., does $Person.age \times Person.height$ yield anything meaningful?

✓ What about $LineItem.quantity \times LineItem.ItemPrice$?

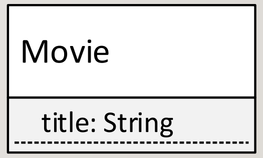
ATTRIBUTE VERSUS RELATIONSHIP





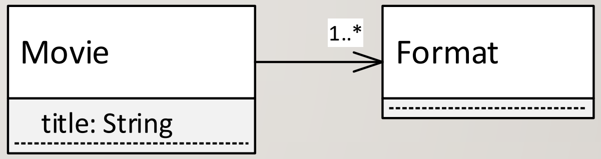
THE STAR TREK PROBLEM

- “Star Trek II: The Wrath of Khan is a movie”



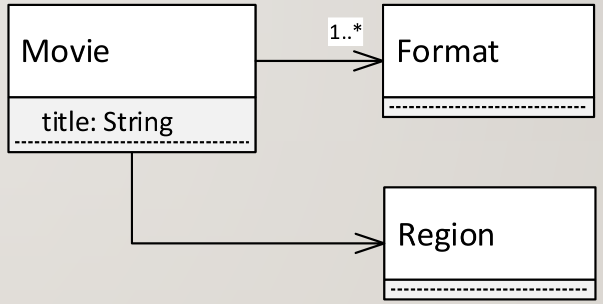
THE STAR TREK PROBLEM

- “Star Trek II: The Wrath of Khan is a movie”
- It is available on Blu-ray and DVD



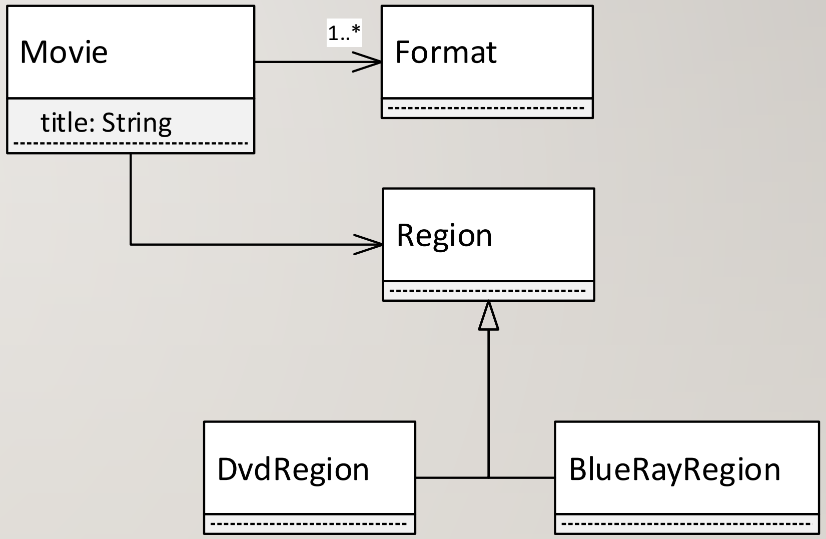
THE STAR TREK PROBLEM

- “Star Trek II: The Wrath of Khan is a movie”
- It is available on Blu-ray and DVD
- DVDs are region-coded



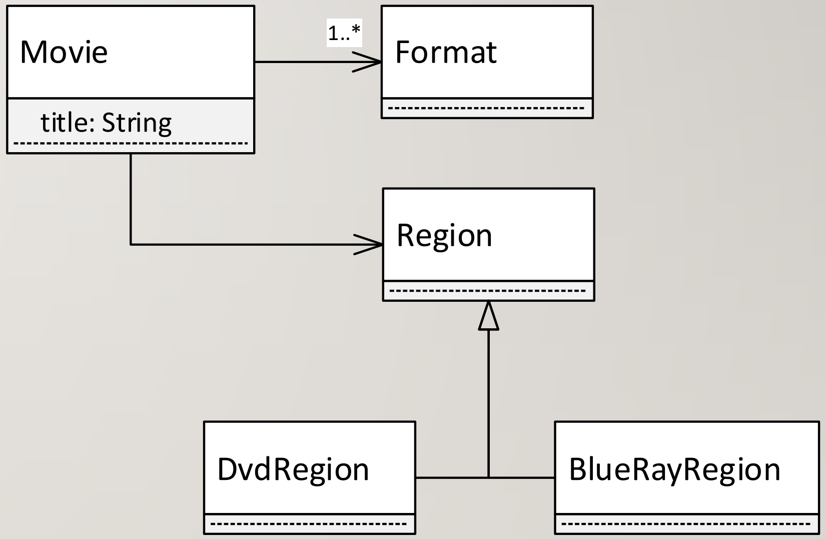
THE STAR TREK PROBLEM

- “Star Trek II: The Wrath of Khan is a movie”
- It is available on Blu-ray and DVD
- DVDs are region-coded
- Blu-rays are region-coded with different regions



THE STAR TREK PROBLEM

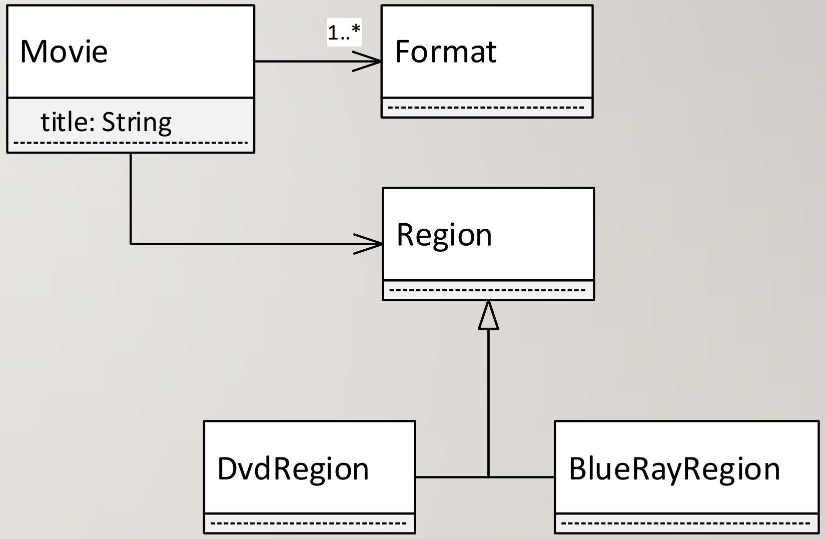
- "Star Trek II: The Wrath of Khan is a movie"
- It is available on Blu-ray and DVD
- DVDs are region-coded
- Blu-rays are region-coded with different regions



THE STAR TREK PROBLEM

- “Star Trek II: The Wrath of Khan is a movie”
- It is available on Blu-ray and DVD
- DVDs are region-coded
- Blu-rays are region-coded with different regions

- “I want all Star Trek movies with the original cast.”
- “I want all Star Trek movies not in the JJ Abrams timeline.”
- Now track inventory of each format of disc in each store location (entity == type of disc)
- Now track digital licenses issues for online viewing (entity == individual playback)



THE STAR TREK PROBLEM

- Precise, normalized models exist to exclude invalid values.
- Hard to use when previously unknown or invalid dimensions come into play.
- Grouping or collecting in new ways requires new attributes,
  which means more precision in the model.

TAKEAWAYS

- Data is not reality
- We represent parts of reality in our systems
- Models exist to represent data
- But they also exclude whatever can’t be represented

SOME DIFFICULT IDEAS

SOME DIFFICULT IDEAS

- Data duplication is OK. Just make sure you can reconcile the duplicates.

SOME DIFFICULT IDEAS

- Data duplication is OK. Just make sure you can reconcile the duplicates.
- There is no "natural" data model

SOME DIFFICULT IDEAS

- Data duplication is OK. Just make sure you can reconcile the duplicates.
- There is no "natural" data model
- Relational, KVS, hierarchic, network, document, graph… they're all just ways to group attributes into entities.

SOME DIFFICULT IDEAS

- Data duplication is OK. Just make sure you can reconcile the duplicates.
- There is no "natural" data model
- Relational, KVS, hierarchic, network, document, graph… they're all just ways to group attributes into entities.
- Each one is optimized for certain kinds of transaction and query patterns

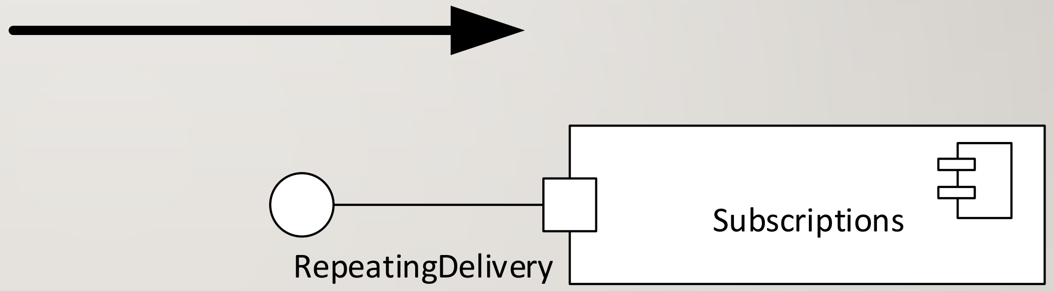
SOME DIFFICULT IDEAS

- Data duplication is OK. Just make sure you can reconcile the duplicates.
- There is no "natural" data model
- Relational, KVS, hierarchic, network, document, graph… they're all just ways to group attributes into entities.
- Each one is optimized for certain kinds of transaction and query patterns
- Too much normalization impedes future adaptability

IDENTIFIERS AND OWNERSHIP

STUDY IN IDENTIFIERS

```
{
  “item”: “123123123”,
  “party”: “99349394”,
  “scheduleType”: “1”
}
```



Suppose we need details about “123123123”
How do we know what system to call?

“NAKED” IDENTIFIERS HAVE IMPLICIT CONTEXT

Must have code that knows how to get details:

- Host
- Port
- Protocol
- Type of query
- Format of results

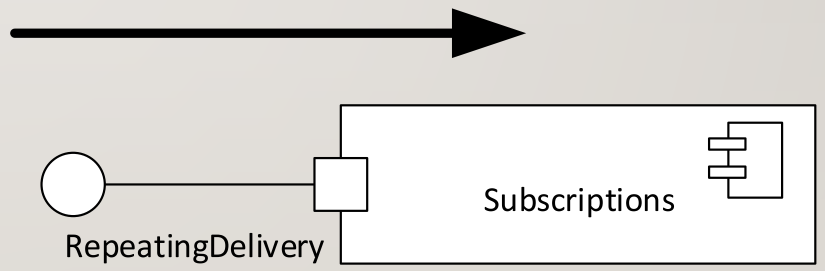
Also assumes just one authority for details

USE EXPLICIT CONTEXT

- Make context explicit to allow multiple authorities: URL or URN
- Enable many sources
- Standardize the media type or representation
- Use logical names, not application or service names
- Proxies and rewrite rules can keep the URLs functioning

BETTER

```
{
  “item”: “https://example.com/skus/123123123”,
  “party”: “https://example.com/people/99349394”,
  “schedule”: “schedule:weekly”
}
```



REQUIRED FIELDS

REQUIRED FOR WHAT?

- Recall: data model determines what to exclude
- Changing requirements harder to support with more restrictions on data

EXAMPLE: PROX CARD TABLE FOR BIKE SHARING

Field	Type	Constraint
User_id	String	Not null
Card_type	String	Not null, enumerated
Status	String	Not null, enumerated
Name	String	Not null

For each "not null," can you think of a valid use (or future change) that it disallows?

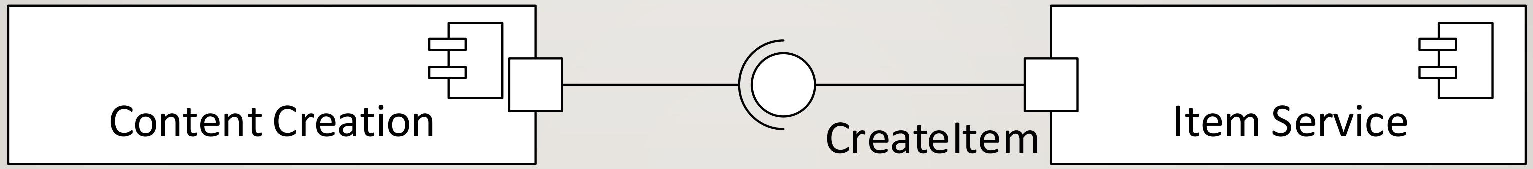
“IS VALID” METHODS

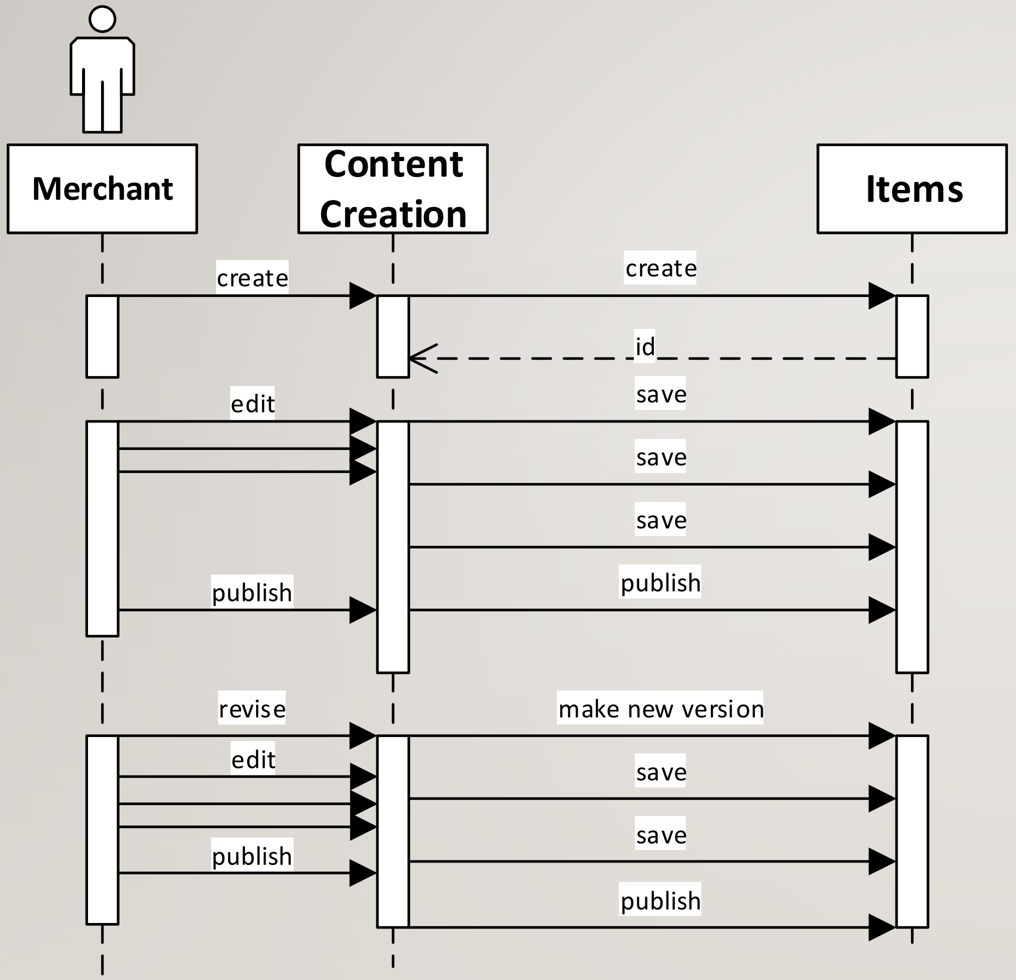
- Again: “is valid” for what operation?
- “Valid” implies policy
- Recall: Contextualize downstream

EXAMPLE: INSIDE A COMMERCE SYSTEM

```
public interface Item {
  String getName(int version);
  void setName(int version, String name);

  String getDescription(int version);
  void setShortDescription(int version, String shortDescription);

  boolean isValid(int version);

  void publishItem(int version);
  int getLatestVersion();
  int[] getAllVersions();
  …
}
```

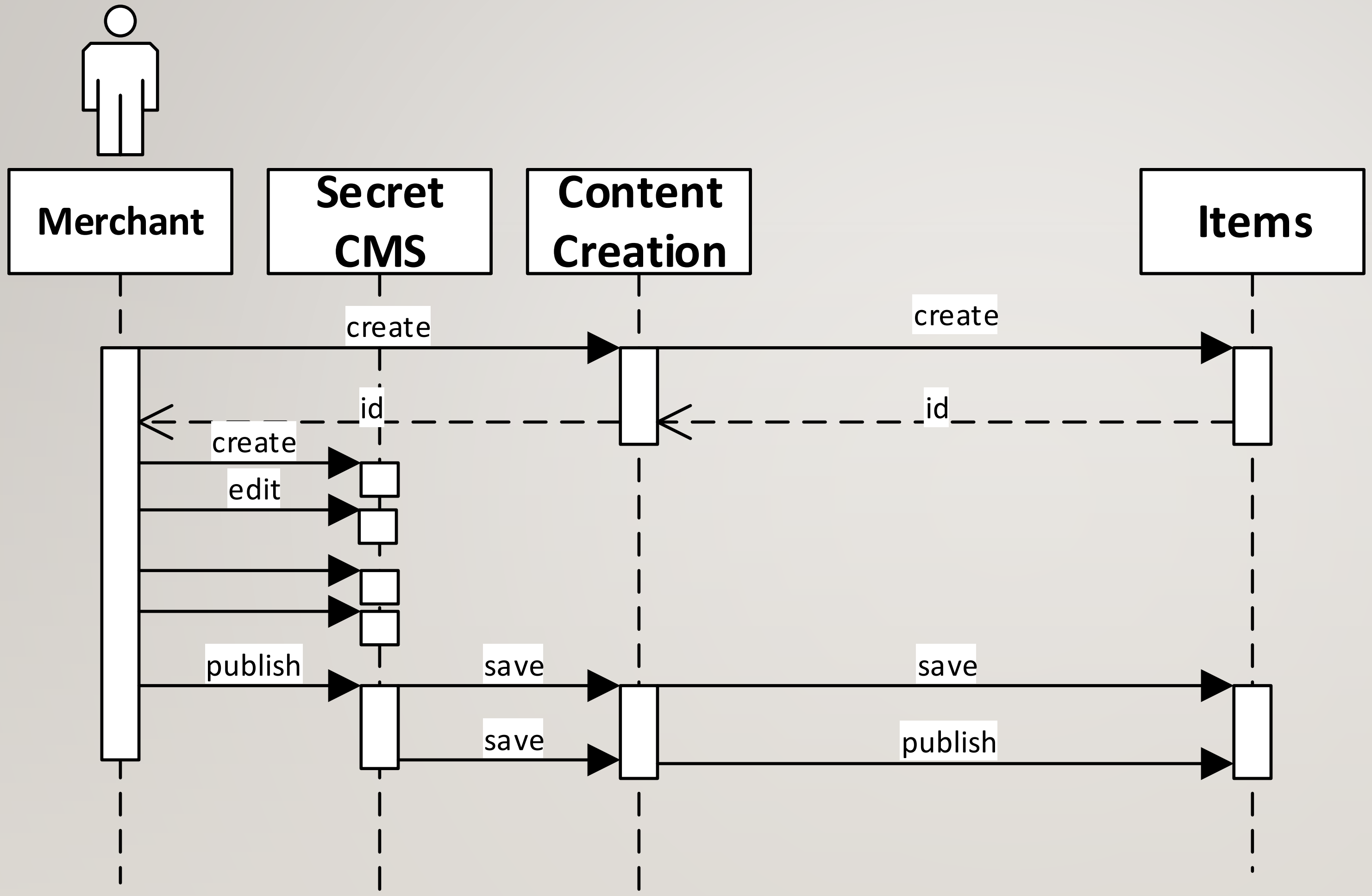
EXAMPLE: INSIDE A COMMERCE SYSTEM





EXAMPLE: EXPECTED INTERACTION

WHAT REALLY HAPPENED



INTERFACE SEGREGATION

```
public interface Item {
  String getName(int version);
  void setName(int version, String name);

  String getDescription(int version);
  void setShortDescription(int version, String shortDescription);

  boolean isValid(int version);

  void publishItem(int version);
  int getLatestVersion();
  int[] getAllVersions();
  …
}
```

❌

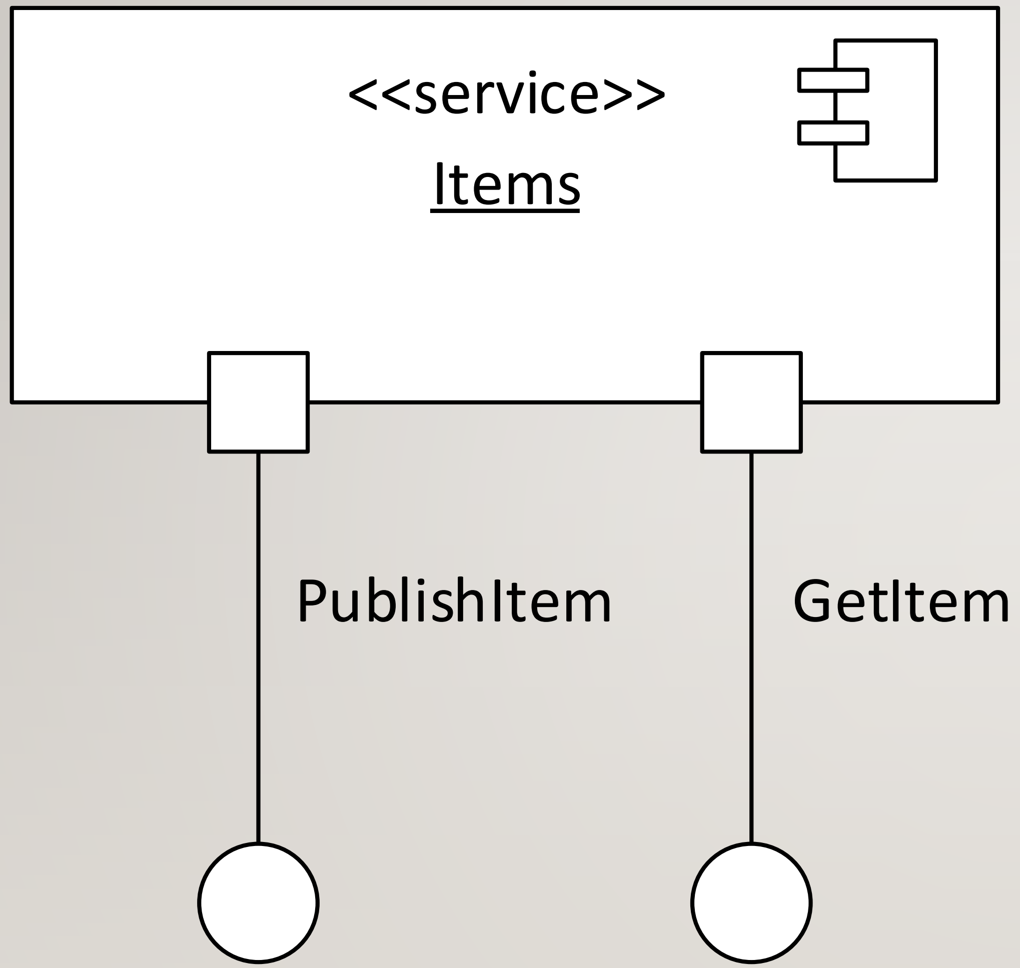
INTERFACE SEGREGATION – IN CODE

```
public interface Versions<T> {
  <T> getLatestVersion();
  List<T> getAllVersions();
  void publish(T details);
  ...
}
```

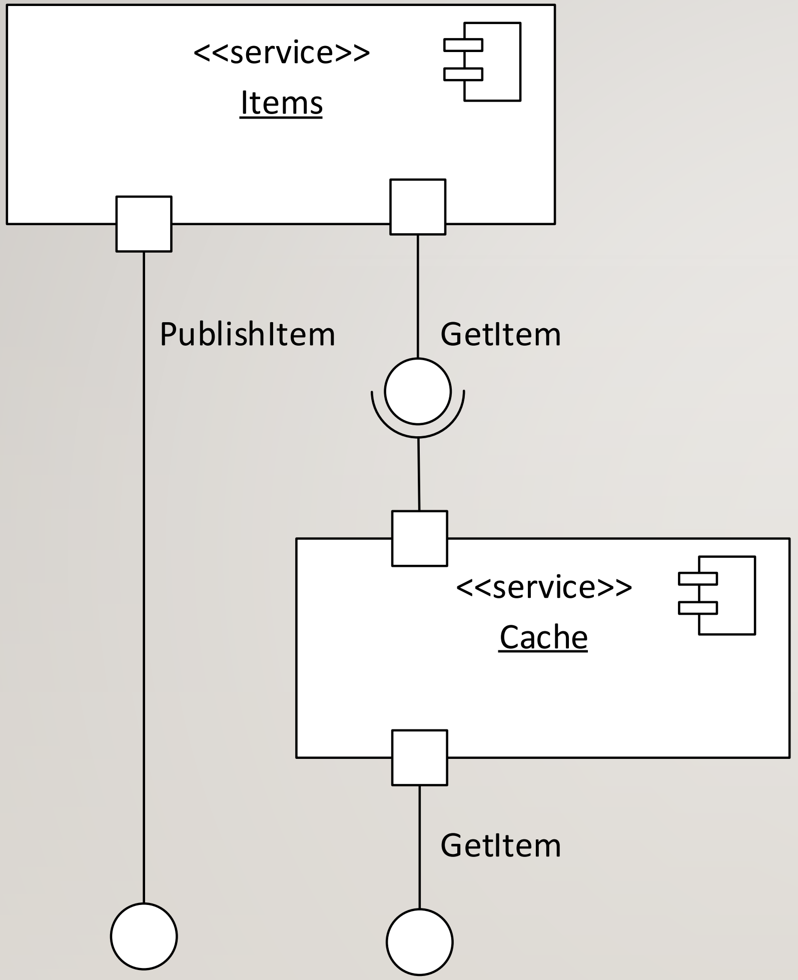
```
public interface Item {
  String getName();
  String getShortDescription();
  ...
}
```

✓

INTERFACE SEGREGATION – AT SERVICE LEVEL



- Caller uses the interface it needs
- Caller isn't aware both are from same component



INTERFACE SEGREGATION – AT SERVICE LEVEL

- Caller uses the interface it needs
- Caller isn't aware both are from same component
- Allows intermediation or substitution

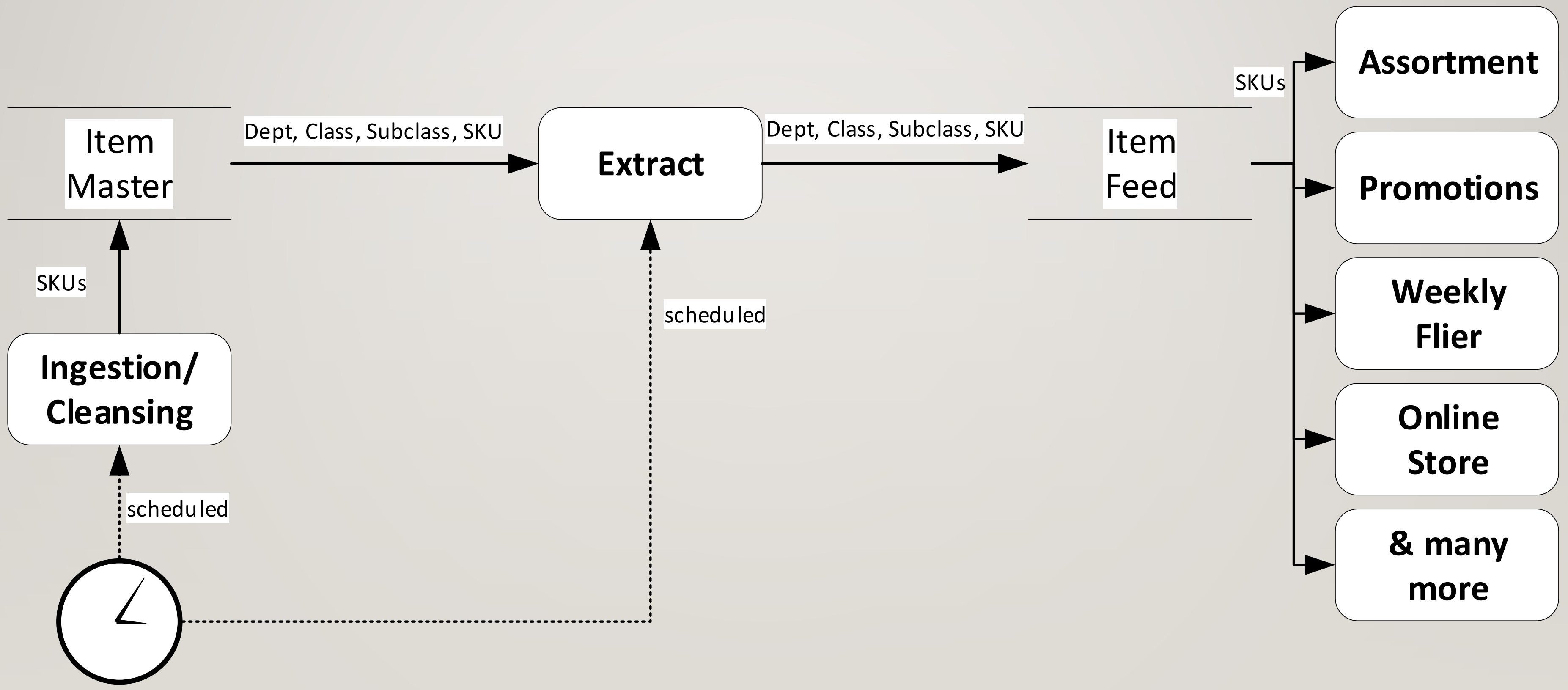
SPLITTING NOUNS

DON'T GET FOOLED AGAIN

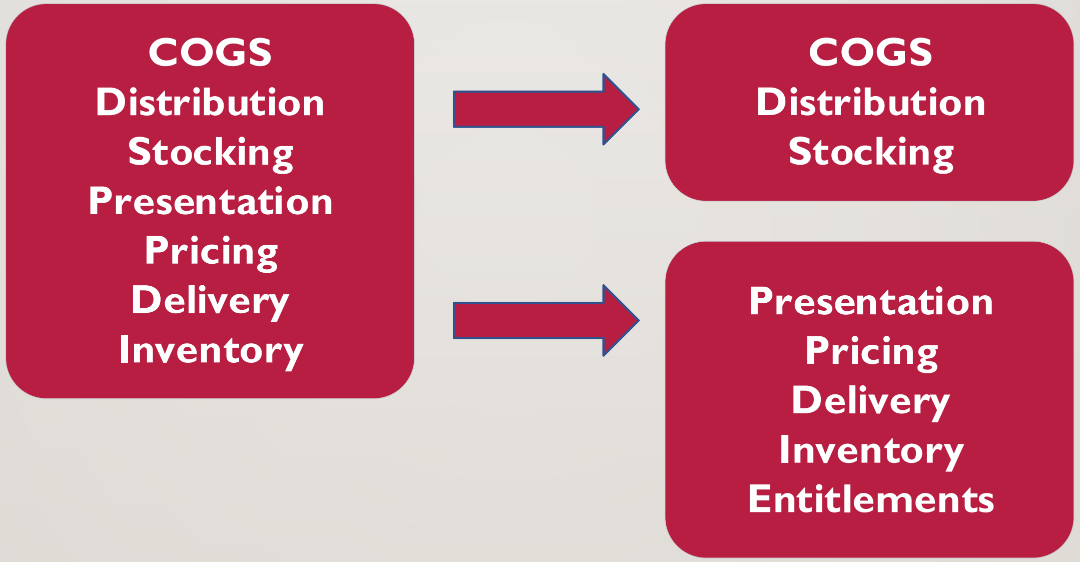
- People thought of "SKU" as a real thing, forgot that it's an abstraction
- Many attributes needed for diverse purposes.

COGS
Distribution
Stocking
Presentation
Pricing
Delivery
Inventory

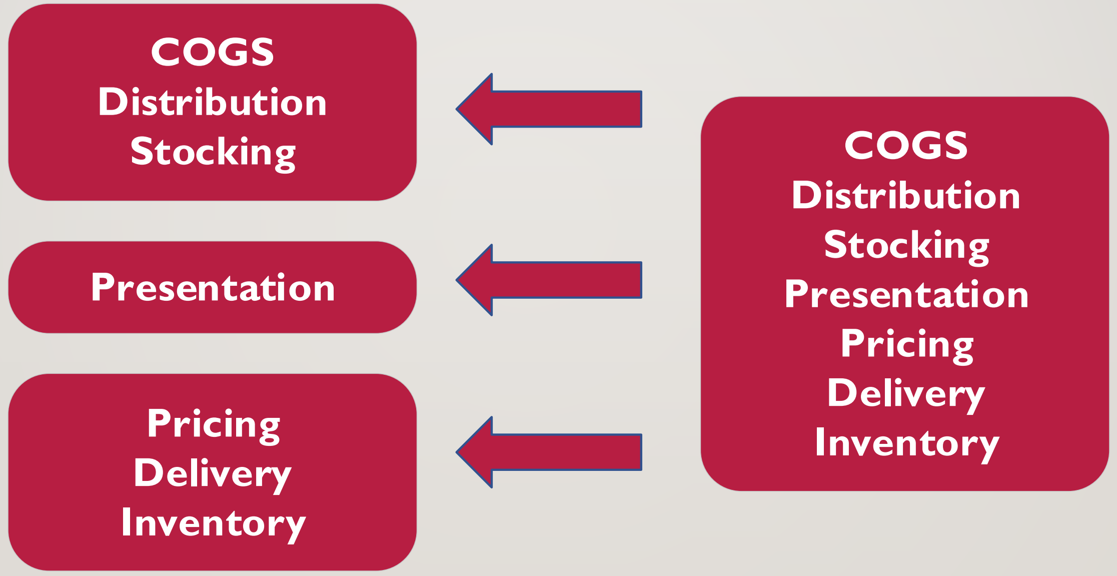
"ITEM MASTER"


Item Master
Dept, Class, Subclass, SKU
Extract
Dept, Class, Subclass, SKU
Item Feed
SKUs
Assortment
Promotions
Weekly Flier
Online Store
& many more
SKUs
scheduled
Ingestion/ Cleansing
scheduled

FOR SELLING DIGITAL GOODS



FOR MARKETPLACE



BETTER

- Use SKU *strictly* as an identifier (preferably with a URL)
- Offer interfaces to exchange a SKU for display attributes, shipping attributes, etc.
- Allow pre-packaged aggregates: sku-delivery-shipping, sku-pricing, etc.
- Decouple systems from the parts of SKU they don't need.
    - E.g., warehouse doesn't care about 27 different images of the product.

CHALLENGE FOR MICROSERVICES

- Instinct: Create “Noun” based services
- Leads to the “Entity Service” antipattern

http://www.michaelnygard.com/blog/2017/12/the-entity-service-antipattern/

http://www.michaelnygard.com/blog/2018/01/services-by-lifecycle/

ACTIVITY:
MODEL THE STUFF SUBSCRIBERS CAN GET

- Might be something delivered to their home or business
- Might be a service that someone performs for them (e.g., dog walking)
- Requires recurring payment
- Will be delivered periodically, maybe on different schedule than payment
- Should be attractive and help convert browsers to subscribers

ACTIVITY: MODEL THE STUFF SUBSCRIBERS CAN GET

1. Start listing all the attributes you can think of.
2. Now think about common use cases:
   What does a subscriber need to see when browsing?
   What does a vendor need to set up?
   What about a CSR handling an upset customer call?
3. Group your attributes by use case.
4. Now, finally, think of good names for those groups.